Visualizing metagenomic data in R using Jetstream

HALEY LEFFLER, Indiana University
SHERI A. SANDERS, National Center for Genome Analysis Support
BHAVYA PAPUDESHI, National Center for Genome Analysis Support

CCS Concepts: • **Applied computing** → *Bioinformatics.*

Additional Key Words and Phrases: datasets, visualization, metagenomes

ACM Reference Format:
Haley Leffler, Sheri A. Sanders, and Bhavya Papudeshi. 2020. Visualizing metagenomic data in R using Jetstream. 1, 1 (August 2020), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Microbial communities play an integral role in maintaining human and ecosystem health. Metagenomics is a culture-independent technique allowing researchers to study microbial communities from samples collected directly from the environment [1, 2, 6]. Researchers applying metagenomics techniques generate large datasets to capture the complexity of the microbial community and face challenges during visualization [1, 2]. Using several visualization methods to display the data from multiple samples allows for exploratory analysis and represents the data from different points of view [6].

Test data for this project was downloaded from a research study focusing on the microbial response to hydrocarbon seepages (the natural release of oil or gas bubbles from the ocean floor). This study found that hydrocarbon seepages can significantly alter the microbial community [8]. Most of the reads in this dataset remain unidentified due to the sampled environment's novelty, making it an interesting subject for visualization. Generally, because datasets only capture a subset of the microbial community, one must begin by asking if it is a representative sample. To address this, a rarefaction curve can be applied. A representative sample would show an exponential increase as new species are identified, plateauing as unique species are identified. The other visualization method to plot datasets with a large number of variables are ordination plots such as Principal Component Analysis (PCA) and non-metric Multidimensional Scaling (nMDS). These plots visualize patterns or gradients, inferring similarity between samples and spotting anomalies. Additionally, heatmaps explore the differences between the samples replacing abundance with a color scale. The data can also be visualized using alluvial plots to explore differences between samples based on their metadata [4]. This project performs an exploratory analysis of metagenomes using Jetstream, a cloud-based infrastructure that aids analysis of large datasets [5].

Authors' addresses: Haley Leffler, Indiana University, hleffler@iu.edu; Sheri A. Sanders, National Center for Genome Analysis Support, ss93@iu.edu; Bhavya Papudeshi, National Center for Genome Analysis Support, bhnala@iu.edu.

XXXX-XXXX/2020/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 METHODS

2.1 Materials and Data Collection

The metagenomic dataset collected from BioProject number PRJNA553005 contains seven sediment samples from the Gulf of Mexico. The dataset was downloaded from the Sequence Read Archive (SRA) using sratoolkit. The samples were collected from three sites: site 1 (seep1 - D24, D27, and D23), site 2 (seep2 - D72 and D75), and reference (D21 and D30). D24 contained characteristics from all three sites and was labeled as transition sample. The dataset was analyzed on an Ubuntu base Jetstream virtual machine (VM) (CPU: 6, Memory: 16 GB, Disk: 60 GB) with Jupyter notebook (5.2.2) and R (4.0.2) installed.

2.2 Taxonomic annotation

Using the Kraken (2.0.8) toolkit [7], samples were aligned against the microbial database and given taxonomic reports. Reports were combined into a table for R compatibility, and were used as input for the visualizations.

2.3 Visualization Methods

We used different inputs for each method to explore differences in taxonomic ranks among the samples. First, we used the taxonomic report with family-level classifications and their corresponding abundance as input for rarefaction curves and ordination plots (PCA, nMDS) using vegan package. Next, for heatmaps, species-level classification and their corresponding abundance(%) was plotted using the ComplexHeatmap package from Bioconductor v3.11. Finally, phylum-level classification with their corresponding abundance(%) was used as input for alluvial plots using ggalluvial and ggplot2 packages.

3 RESULTS

Kraken taxonomic reports revealed that 88-90% of the sequences remain unidentified per sample. Regardless of the level of classification, identified taxa have low (<9%) abundance in total. From the rarefaction curve, we were able to determine that the metagenomes collected were representative as the samples begin to plateau when 475 unique species were identified per 1.5 million sequences. In the two ordination plots, the seep samples and reference samples cluster together respectively, showing higher correlations of the microbial profiles between them. The heatmap showed that *Homo sapiens* is most abundant, and there are species-level differences between seep and reference sites. Looking further into these taxonomic changes in the alluvial plot, the transition sample (D24) has a similar taxonomic profile to both the reference and the seep samples.

4 DISCUSSION

After exploratory analysis of the hydrocarbon seepage datasets, our conclusions align with Zhao *et al.*,[8] and reveal more about the samples' representativeness and contamination. The rarefaction curve confirmed that the dataset is representative of the larger population. The ordination plots revealed that samples collected from seep sites are more similar compared to samples from reference sites. The heatmap showed *Homo sapiens* in greatest abundance, suggesting human DNA contamination and showing the necessity of quality control before further analysis. The alluvial plot highlights the similarities of D24 samples against both the seep and reference locations. These results propose that the availability of nutrients in the sites may be driving change in microbial taxa and influencing their functional profiles. While the visualization methods perform exploratory analysis, each method's limitations should be considered, and no statistical analysis was done to

show significance. The scripts written to make the plots and input data are available on GitHub in a Jupyter notebook for public use[3].

REFERENCES

[1] Florian P Breitwieser, Jennifer Lu, and Steven L Salzberg. 2019. A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics* 20, 4 (2019), 1125–1136.

[2] Philip Hugenholtz and Gene W Tyson. 2008. Metagenomics. *Nature* 455, 7212 (2008), 481–483.

[3] Haley Leffler and Bhavya Papudeshi. 2020. GitHub REU Microbial Visualization. https://github.com/hleffler/REU-microbial-visualization

[4] Martin Rosvall and Carl T Bergstrom. 2010. Mapping change in large networks. *PloS one* 5, 1 (2010), e8694.

[5] Craig A Stewart, Timothy M Cockerill, Ian Foster, David Hancock, Nirav Merchant, Edwin Skidmore, Daniel Stanzione, James Taylor, Steven Tuecke, George Turner, et al. 2015. Jetstream: a self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. 1–8.

[6] Konstantin Sudarikov, Alexander Tyakht, and Dmitry Alexeev. 2017. Methods for the metagenomic data visualization and analysis. *Curr Issues Mol Biol* 24 (2017), 37–58.

[7] Derrick E Wood, Jennifer Lu, and Ben Langmead. 2019. Improved metagenomic analysis with Kraken 2. *Genome biology* 20, 1 (2019), 257.

[8] Rui Zhao, Zarath M Summers, Glenn D Christman, Kristin M Yoshimura, and Jennifer F Biddle. 2020. Metagenomic views of microbial dynamics influenced by hydrocarbon seepage in sediments of the Gulf of Mexico. *Scientific reports* 10, 1 (2020), 1–13.